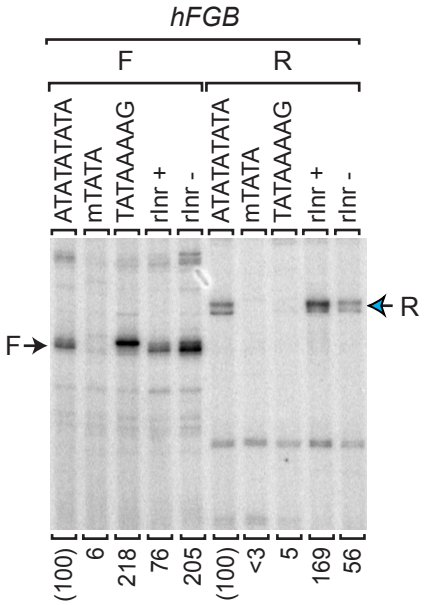
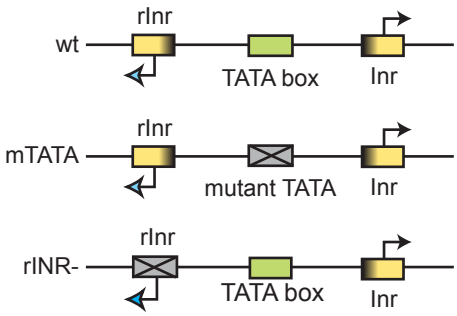
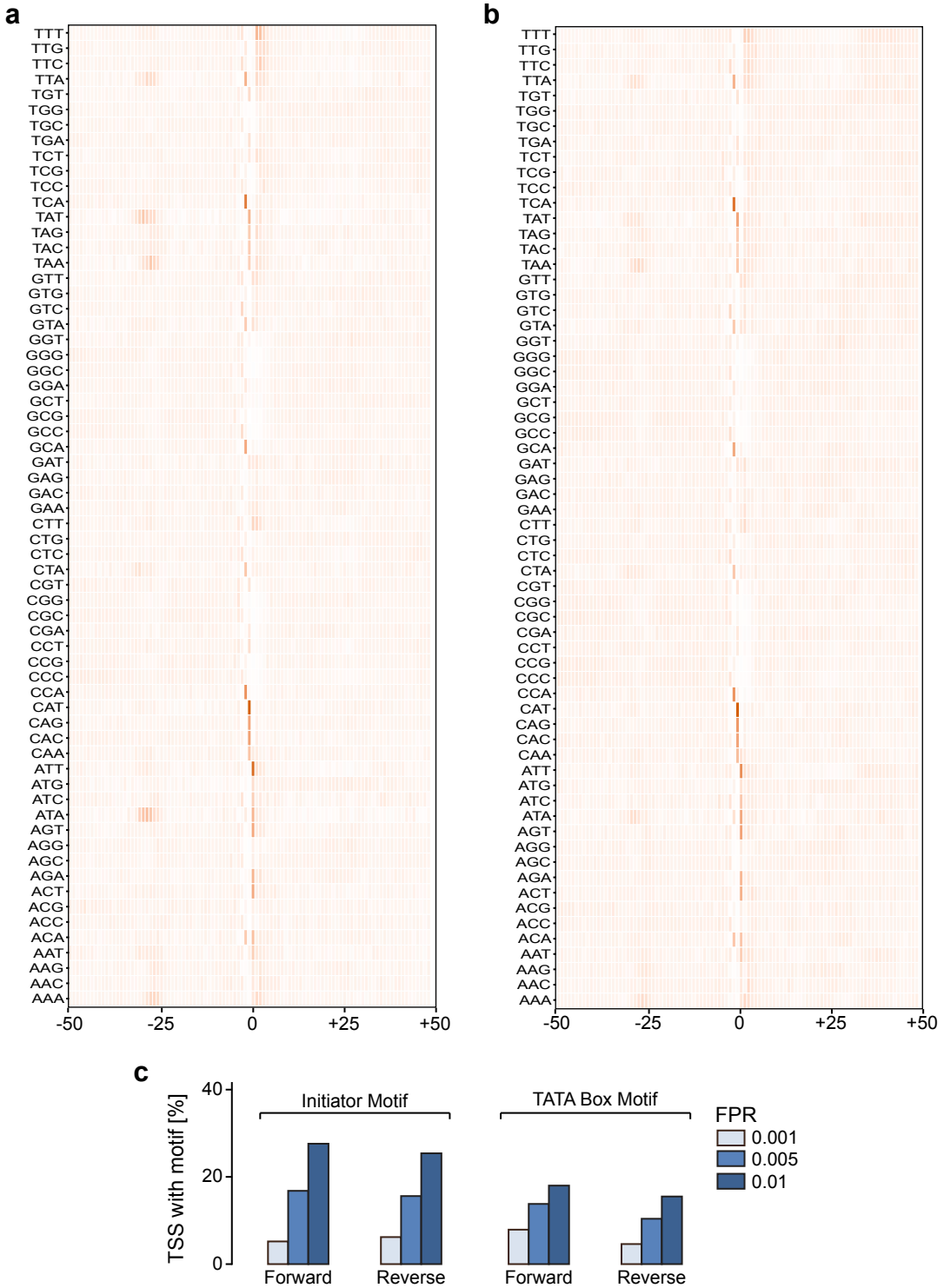
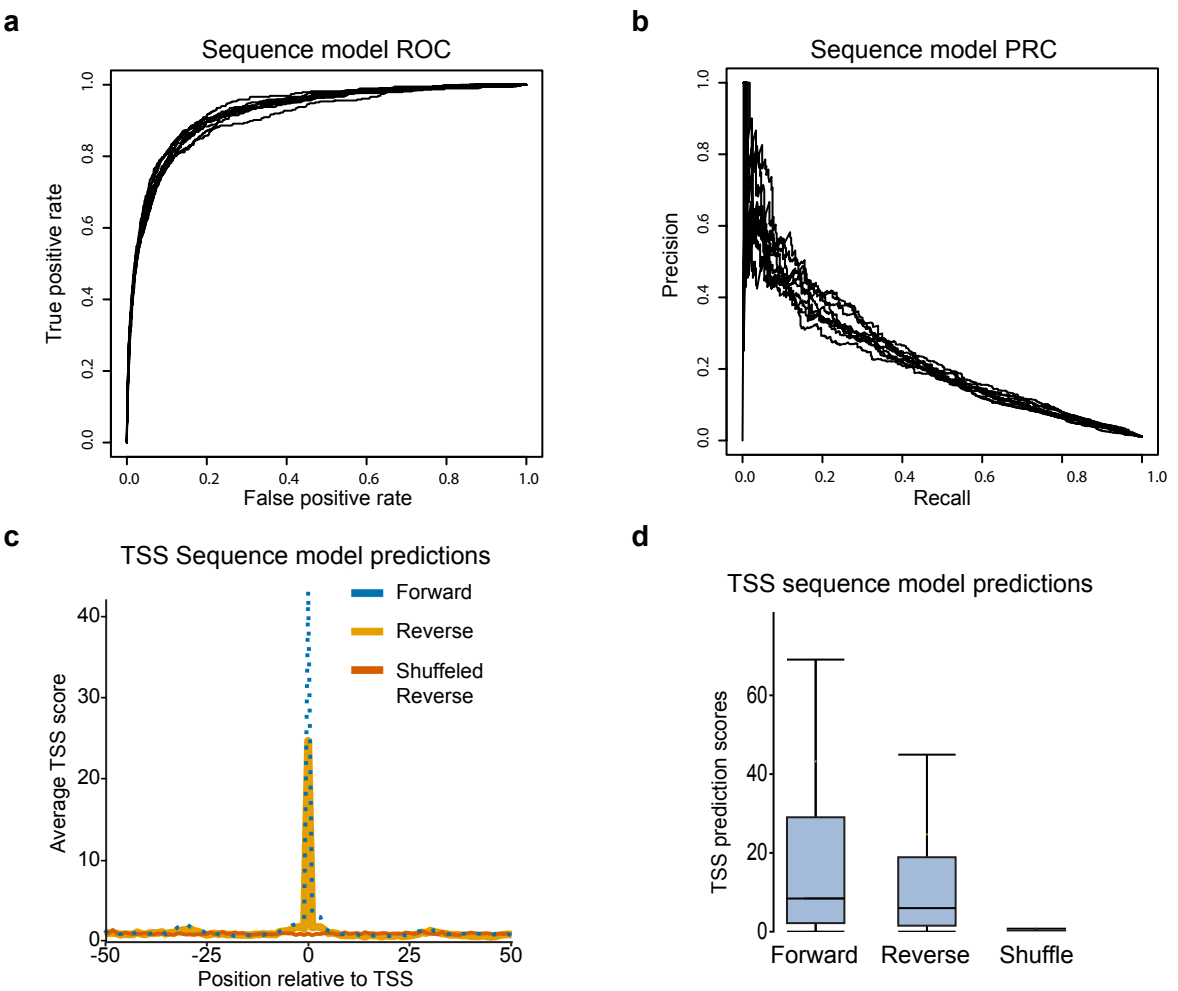
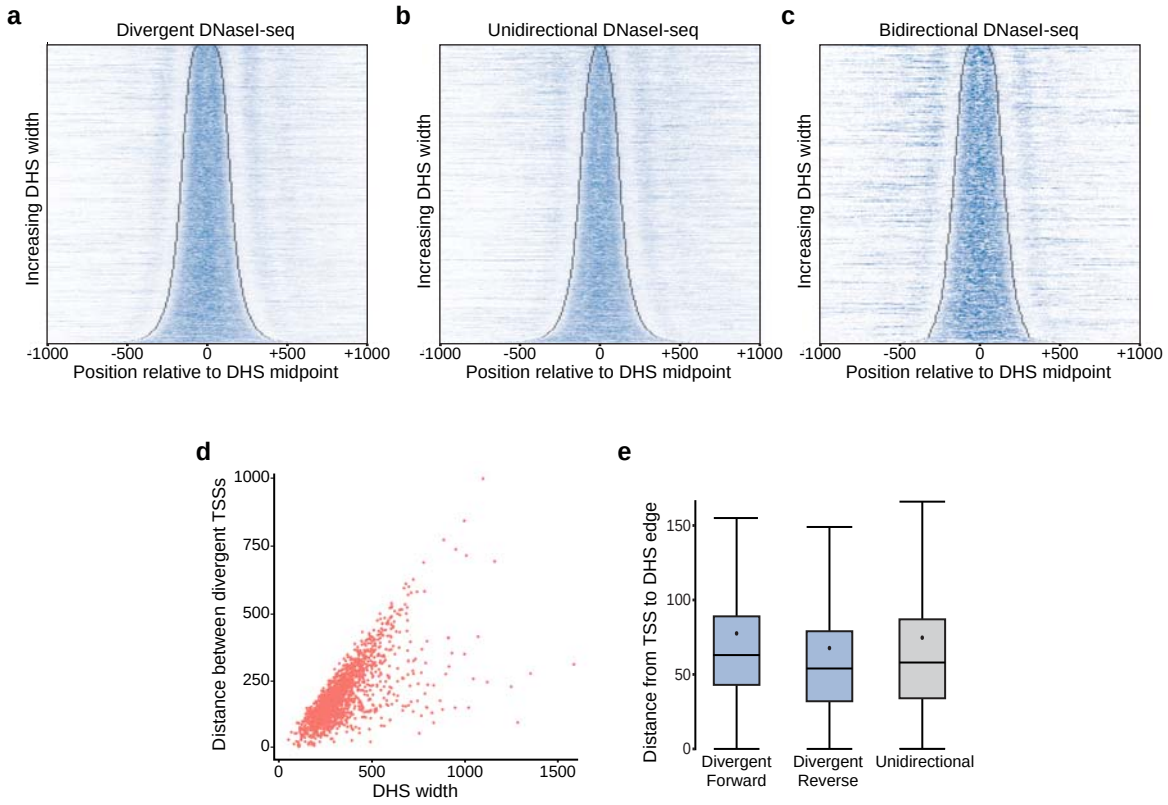

hFGB
F
R
ATATATATA
mTATA
TATAAAAG
rInr +
rInr -
ATATATATA
mTATA
TATAAAAG
rInr +
rInr -
F
R
(100)
6
218
76
205
(100)
<3
5
169
56

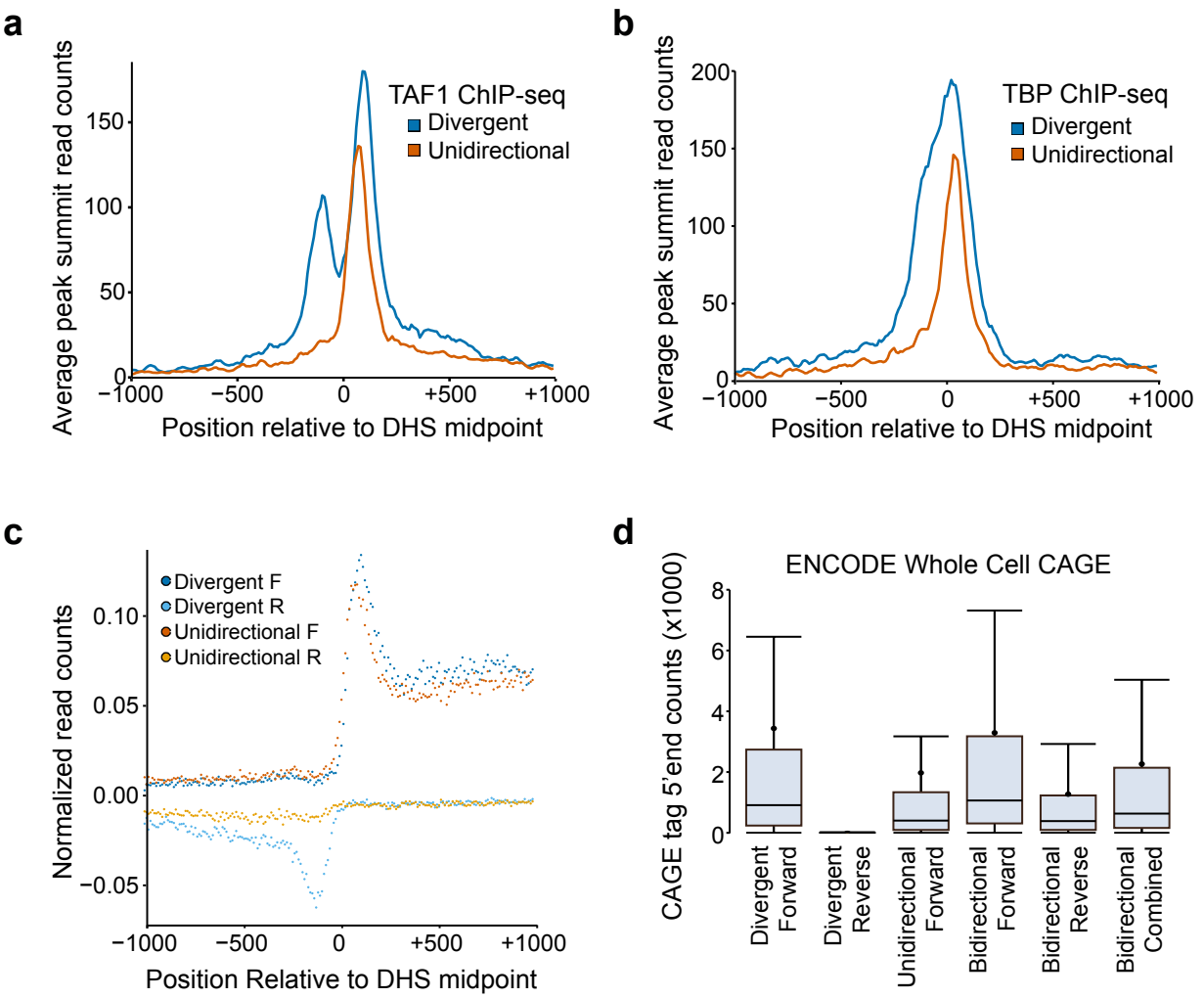

wt
rInr
TATA box
Inr
mTATA
rInr
mutant TATA
Inr
rINR-
rInr
TATA box
Inr


a
b
TTT
TTG
TTC
TTA
TGT
TGG
TGC
TGA
TCT
TCG
TCC
TCA
TAT
TAG
TAC
TAA
GTT
GTG
GTC
GTA
GGT
GGG
GGC
GGA
GCT
GCG
GCC
GCA
GAT
GAG
GAC
GAA
CTT
CTG
CTC
CTA
CGT
CGG
CGC
CGA
CCT
CCG
CCC
CCA
CAT
CAG
CAC
CAA
ATT
ATG
ATC
ATA
AGT
AGG
AGC
AGA
ACT
ACG
ACC
ACA
AAT
AAG
AAC
AAA
-50
-25
0
+25
+50
c
TSS with motif [%]
40
20
0
Initiator Motif
TATA Box Motif
FPR
0.001
0.005
0.01
Forward
Reverse
Forward
Reverse


a
Sequence model ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
b
Sequence model PRC
Precision
Recall
c
TSS Sequence model predictions
Forward
Reverse
Shuffeled Reverse
Average TSS score
Position relative to TSS
-50
-25
0
25
50
10
20
30
40
d
TSS sequence model predictions
TSS prediction scores
Forward
Reverse
Shuffle
0
20
40
60

Supp. Figure-4 (Ohler)




a
Average peak summit read counts
150
100
50
0
TAF1 ChIP-seq
Divergent
Unidirectional
−1000
−500
0
+500
+1000
Position relative to DHS midpoint
b
Average peak summit read counts
200
150
100
50
0
TBP ChIP-seq
Divergent
Unidirectional
−1000
−500
0
+500
+1000
Position relative to DHS midpoint
c
Normalized read counts
0.10
0.05
0.00
−0.05
Divergent F
Divergent R
Unidirectional F
Unidirectional R
−1000
−500
0
+500
+1000
Position Relative to DHS midpoint
d
ENCODE Whole Cell CAGE
CAGE tag 5'end counts (x1000)
8
6
4
2
0
Divergent Forward
Divergent Reverse
Unidirectional Forward
Bidirectional Forward
Bidirectional Reverse
Bidirectional Combined

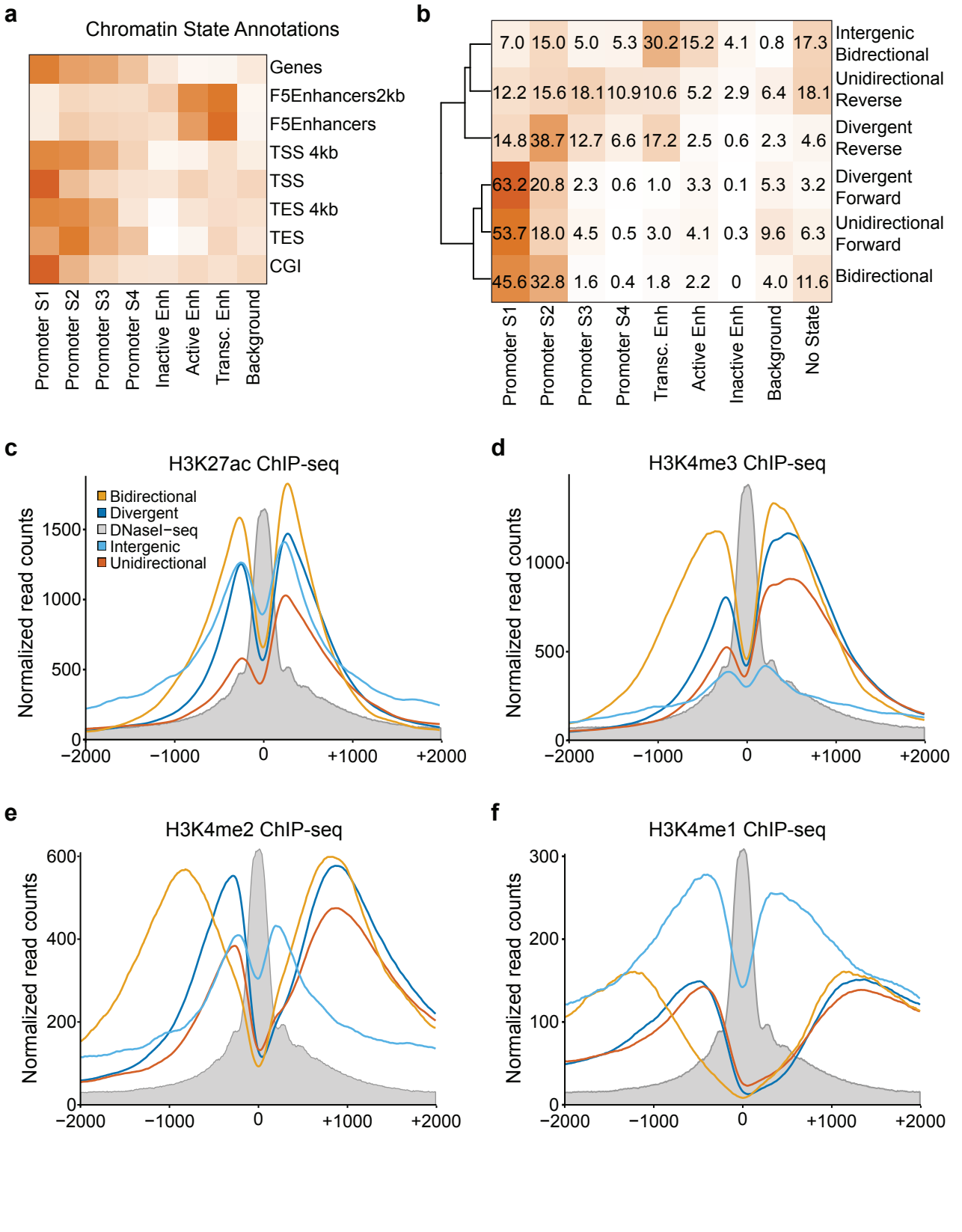

a
Chromatin State Annotations
Genes
F5Enhancers2kb
F5Enhancers
TSS 4kb
TSS
TES 4kb
TES
CGI
Promoter S1
Promoter S2
Promoter S3
Promoter S4
Inactive Enh
Active Enh
Transc. Enh
Background
b
7.0 15.0 5.0 5.3 30.2 15.2 4.1 0.8 17.3
Intergenic
Bidirectional
12.2 15.6 18.1 10.9 10.6 5.2 2.9 6.4 18.1
Unidirectional
Reverse
14.8 38.7 12.7 6.6 17.2 2.5 0.6 2.3 4.6
Divergent
Reverse
63.2 20.8 2.3 0.6 1.0 3.3 0.1 5.3 3.2
Divergent
Forward
53.7 18.0 4.5 0.5 3.0 4.1 0.3 9.6 6.3
Unidirectional
Forward
45.6 32.8 1.6 0.4 1.8 2.2 0 4.0 11.6
Bidirectional
Promoter S1
Promoter S2
Promoter S3
Promoter S4
Transc. Enh
Active Enh
Inactive Enh
Background
No State
c
H3K27ac ChIP-seq
Bidirectional
Divergent
DNaseI−seq
Intergenic
Unidirectional
Normalized read counts
0
500
1000
1500
−2000
−1000
0
+1000
+2000
d
H3K4me3 ChIP-seq
Normalized read counts
0
500
1000
−2000
−1000
0
+1000
+2000
e
H3K4me2 ChIP-seq
Normalized read counts
0
200
400
600
−2000
−1000
0
+1000
+2000
f
H3K4me1 ChIP-seq
Normalized read counts
0
100
200
300
−2000
−1000
0
+1000
+2000

Figure S1 | Transcription from a divergent core promoter, related to Figure 1. Human FGB as an example of a divergent core promoter. The polarity depends on the DNA sequence. Promoters were cloned from +50 to -125 (relative to the +1 transcription start site) to allow reverse initiation within the natural sequence. The reverse Inr (rInr) sequence "TCAGAA" was substituted with "TCGGTC" (rInr-) or a consensus Inr "TCAGTC"(rInr+).

Figure S2 | Sequence content of forward and reverse TSSs, related to Figure 2. **a,b,** Position-specific threemer counts normalized to total threemer frequencies for forward (a) and reverse (b) direction core promoters -50 to +50 bp around the 5'-GRO-seq cluster modes. **c,** Percent of forward or reverse TSSs that show motif matches to either initiator (left) or TATA-box (right) in the -35 to -25 or -5 to +5 regions, respectively, from the 5'-GRO-seq cluster modes. Different colors represent different false positive rate (FPR) cutoffs.

Figure S3 | Performance and results of TSS sequence model, related to Figures 2 and 4. a,b, Receiver operator characteristic (a) and precision-recall (b) curves for the sequence model described in Frith et al, 2008, trained and tested with a 10-fold cross validation +/- 50 bp around the mode of the forward TSSs from the divergent promoter pairs described in Figure 2 (see Experimental Procedures). **c,** Average predicted TSS scores per position for sequences +/- 50 bp around the mode of the corresponding TSSs from the divergent promoter pairs described in Figure 2, or its shuffled control, from the model trained as in "a" and "b" (see Supplemental Experimental Procedures). **d,** Distributions of 5'GRO-seq cluster mode TSS prediction scores for forward and reverse TSSs.

Figure S4 | DHS peak call accuracy and characteristics, related to Figures 4 and 5. a,b,c, Heat maps of normalized DNaseI-seq read 5'end counts (blue) anchored on each DHS midpoint and ranked by increasing DHS width together with the location of JAMM-called peak edges (black) for divergent (a), unidirectional (b), and bidirectional (c) promoter DHSs. **d,** Scatter plot of DHS width versus distance between forward and reverse 5'-GRO-seq cluster modes of divergent promoters. **e,** Boxplots of distance between 5'-GRO-seq cluster modes and corresponding DHS edges, dot = mean.

Figure S5 | Unidirectional promoters lack upstream hallmarks of divergent transcription, related to Figures 4 and 5. a,b, Positional average fragment-extended ChIP-seq read counts within Taf1 (a) and Tbp (b) peak summits as called by SISSRS in bins of 10 nucleotides (see Supplemental Experimental Procedures). **c,** Positional average of normalized read 5'end counts of traditional GRO-seq for the forward (red and blue) or reverse (orange and light blue) directions of the divergent (red and orange) or unidirectional (blue and light blue) promoters ("normalized counts" refers to 0-to-1 scaling of read counts for every DHS window, see

Supplemental Experimental Procedures). **d,** Distributions of whole HeLa cell, polyA-plus CAGE tag 5'end counts from ENCODE intersecting designated 5'GRO-seq clusters.

Figure S6 | Histone modifications HMM characteristics and analysis, related to Figure 6. a, Chromatin state – Genome Annotation enrichment map (see Supplemental Experimental Procedures). "Genes" are entire UCSC gene lengths, "TSS" are UCSC known gene transcription start sites, "TES" are UCSC known gene transcription end sites, "TSS 4kb" and "TES 4kb" are windows centered around UCSC TSSs and TESs respectively going 2kb upstream and downstream, "F5 Enhancers" are enhancers identified by the Fantom5 consortium for the hg19 genome build, "F5 Enhancers 2k" are windows centered around the midpoints of F5 Enhancers going 1kb downstream and 1kb upstream, "CGI" are UCSC "CpG" islands. **b,** Percentage of chromatin state intersections at 75 bp downstream of the NFR edges. "No State" refers to those locations that did not intersect any chromatin state. **c,d,e,f** Average fragment-extended read counts of H3K27ac (c), H3K4me3 (d), H3K4me2 (e), and H3K4me1(f) ChIP-seq in bins of 10 nucleotides for divergent (blue), unidirectional (red), bidirectional (green), and intergenic (light blue) 5'-GRO-seq-containing DHSs (see Experimental Procedures). grey = average DNaseI-seq read 5'end counts for DHSs from all four groups combined.

Table S1 | Comparison of 5'GRO-seq and exosome KD CAGE analyses, related to Figure 4. The same analyses were performed on both datasets using the same DHS peaks calls as described in the Supplemental Experimental Procedures. Margin numbers indicate the number of DHSs that were identified in each group from each dataset. Table numbers indicate the overlap between DHS classes between the two datasets. The most conservative estimate for percentage of unidirectional promoters is 34% (1196/3499) when only considering DHSs with forward gene evidence in both datasets, from which unidirectional DHSs are consistently classified in both datasets and divergent/bidirectional DHSs identified in at least one dataset. It is likely that many of the forward TSS-containing DHSs (unidirectional, divergent, or bidirectional) identified in only one of the two datasets are true; when these are included, we estimate that the true percentage of unidirectional promoters is closer to 44% (3394/7707).

Table S2 | Correlations between 5'GRO-seq and TSS prediction score or H3K27ac ChIP-seq, related to Figure 4.
Spearman Rho correlation values are shown with corresponding p values between 5'GRO-seq read 5'end counts within called clusters (top) and either the TSS prediction score (left top) or H3K27ac ChIP-seq fragment-extended read counts intersecting a window 148 bp downstream of the appropriate DHS peak edge (left bottom).

Table S3 | Final_Cluster Sets.xlsx, related to Figure 1.
5'GRO-seq cluster calls as identified using the strategy described in Ni *et al.* (Ni et al., 2010) and Supplemental Experimental Procedures.

Supplementary Tables

Table S1.

		5'-GRO-seq (n = 4378)		
		Divergent (1741)	Unidirectional (2237)	Bidirectional (400)
Exosome KD CAGE (n = 6828)	Divergent (2890)	1134	490	4
	Unidirectional (3188)	343	1196	1
	Bidirectional (750)	0	1	330

Table S2.

		5'GRO-seq Cluster Read Counts	
		Forward	Reverse
Sequence Model	Forward	0.22 ($p < 0.0001$)	-0.026 ($p = 0.29$)
	Reverse	-0.04 ($p = 0.096$)	0.16 ($p < 0.0001$)
H3K27ac	Forward	0.39 ($p < 0.0001$)	0.0001 ($p = 0.09$)
	Reverse	0.04 ($p = 0.09$)	0.25 ($p < 0.0001$)

Supplemental Experimental Procedures

Cell culture conditions

HeLa S3 cells were grown at 37°C in DMEM (Cellgro) supplemented with 10% FBS (Gibco), 50 U Penicillin and 50 µg Streptomycin per mL (Gibco).

In vitro transcription assays

Core promoter sequences, ±50 bp in respect to the +1 TSS, were cloned into pUC119 (F/F_R) or pUC118 (R) containing a Pol III specific terminator (Duttke, 2014) using XbaI and PstI. A spacer was further inserted into pUC118 to match the distance of the XbaI and PstI cloning sites to the reverse M13 primer site of pUC119. When indicated, the TATA-box was substituted with "ACGTCCGT" (mTATA).

Transcription reactions were carried out as described previously (Duttke, 2014). Briefly, 7 µL of 13 mg/mL human nuclear extract (HSK) were preincubated with 500 ng DNA template in a total volume of 46 µL with a final concentration of 20 mM HEPES-K^+ (pH 7.6); 50 mM KCl; 6 mM $MgCl_2$; 2.5% (w/v) polyvinyl glycol (compound); 0.5 mM DTT; 3 mM ATP; 0.02 mM EDTA and 2% glycerol at 30°C for 75 minutes. Transcription was started by addition of 4 µL NTPs (5 mM each), carried out for 20 minutes and stopped by addition of 100 µL STOP buffer [20 mM EDTA; 200 mM NaCl; 1% SDS, 0.3 mg/mL glycogen]. After mixing, 12.5 µg Proteinase K was added and reactions were incubated at room temperature (~21°C) for 15 minutes. Nucleic acids were subsequently extracted by standard phenol/chloroform purification followed by ethanol precipitation. Transcripts were subjected to primer extension analysis using 5′- ^{32}P-labeled M13 reverse sequencing primer [5'-AGCGGATAACAATTTCACACAGGA] and separated by urea-

polyacrylamide gel electrophoresis. Gels were exposed to a phosphor imager plate and reverse transcription products visualized and quantified with a Typhoon imager (GE Health Sciences).

5'GRO-seq and GRO-seq library generation and sequencing

5'GRO-seq was performed as described previously (Lam et al., 2013). Briefly, about 10^7 HeLa S3 nuclei were used for run-on with BrU-labelled NTPs. Reactions were stopped by addition of 450 µL Trizol LS reagent (Invitrogen). After RNA extraction and treatment with Turbo DNase (Ambion), both according to the manufacturer's instructions, RNA was hydrolyzed by Zn^{2+} fragmentation (Ambion). The fragmented transcripts were incubated for 2 h at 37°C with polynucleotide kinase (PNK, NEB) at pH 5.5 to remove 3' phosphates. BrU-labelled nascent transcripts were subsequently immunoprecipitated with anti-BrdU agarose beads (Santa Cruz Biotech). For 5'GRO-seq, immunoprecipitated RNA was dephosphorylated with calf intestinal phosphatase (NEB). Then 5′ capped fragments were de-capped with tobacco acid pyrophosphatase (Epicentre). Illumina TruSeq adapters were ligated to the RNA 3′ and 5′ ends with truncated mutant RNA ligase 2 (K227Q) and RNA ligase 1 (NEB), respectively. Reverse transcription was performed with Superscript III (Invitrogen) followed by PCR amplification for 12 cycles. Final libraries were size selected on PAGE/TBE gels to 175–225 bp.

GRO-seq was essentially performed as 5'GRO-seq but the immunoprecipitated RNA was directly de-capped with tobacco acid pyrophosphatase (Epicentre) and subsequently kinased with PNK (NEB) prior to adapter ligation.

5'-GRO-seq and GRO-seq read processing, cluster calls, and annotation

Two replicates of 5'end sequenced reads from the 5'-GRO-seq or traditional GRO-seq protocols were trimmed for adapters using cutadapt (Martin, 2011), mapped together to the hg19 human

genome using Bowtie2 with default settings(Langmead and Salzberg, 2012). Reads that did not map uniquely and reads overlapping rRNA loci were removed, yielding 27,512,149 5'-GRO-seq reads and 21,765,842 traditional GRO-seq reads. Clusters were identified according to the strategy described in Ni *et al.* (Ni et al., 2010). Briefly, a kernel density estimate (KDE) of the 5' end positions of the mapped reads was calculated across the genome. Any region exceeding the genome-wide average KDE that contained at least 10 reads was identified as a cluster and used in subsequent analysis. To annotate the identified clusters, the Genomic Features(Lawrence et al., 2013) R package was used to generate BED files for 5'utr, 3'utr, intron, coding exon, non-coding exon, and promoter (-250 upstream of annotated transcription start sites) regions according to the UCSC knownGenes table. BEDTools (Quinlan and Hall, 2010) intersect was used to perform a prioritized intersection between the 5'-GRO-seq cluster calls and these annotation bed files with the following priorities: transcription start site (TSS) > coding exon > 3'utr > non-coding exon > intron. Clusters intersecting either promoter or 5'utr locations were considered TSS-annotating clusters. Clusters not intersecting any of these locations were considered intergenic-annotating clusters. This strategy resulted in exactly one annotation per 5'GRO-seq cluster. Following downstream analyses of cluster pair calling (either closest upstream or DNase-seq based; described below), regions containing clusters annotated as TSS but that overlapped annotated tRNA loci were removed from subsequent analysis.

DNase-seq and ChIP-seq read processing and peak calling

All 5 datasets of ENCODE-mapped DNase-seq reads for HeLa-S3 cells were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). PCR duplicates from each file were removed using SAMTools (Li et al., 2009). The resulting files were converted to BED using BEDTools (Quinlan and Hall, 2010) and concatenated before peak calling with JAMM v1.0.6

(Ibrahim et al., 2014) (http://code.google.com/p/jamm-peak-finder/, settings: -m narrow -f 1). HeLa-S3 cell, Broad Institute histone modification ChIP-seq raw fastq files were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). Reads were aligned to hg19 genome using Bowtie2 (Langmead and Salzberg, 2012) with default parameters and then filtered for those that did not align uniquely or had more than two mismatches. PCR duplicates were removed after alignment using SAMTools (Li et al., 2009) and converted to standard BED format using BEDTools (Quinlan and Hall, 2010). Histone modification peaks were called using JAMM v1.0.4rev1 (Ibrahim et al., 2014) with default settings while maintaining all replicates separate. The filtered peak lists produced by JAMM were considered for further analysis. Raw ENCODE HeLa-S3 ChIP-Seq fastq files for TAF1 and TBP (Bernstein et al., 2012) were processed in the same way as ENCODE histone modification datasets. Replicate BED files were then concatenated before peaks were called using SISSRS (Narlikar and Jothi, 2012) , which can resolve ChIP-Seq peak summits at high resolution (settings: -s 3095693983).

CAGE read processing

Fastq files from Ntini *et al.*(Ntini et al., 2013) (SRR922110.sra and SRR922111.sra) were obtained from the Gene Expresssion Omnibus (GEO) website. Reads were trimmed according to authors methods (Ntini et al., 2013) using Flexbar (Dodt et al., 2012) and mapped to the hg19 human genome using Bowtie2 with default settings (Langmead and Salzberg, 2012). Reads that did not map uniquely were removed. Mapped .bam files for Hela whole-cell, polyA-plus CAGE were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). CAGE reads were corrected for the 5'end nucleotide bias using the CAGEr R package (http://bioconductor.org/packages/release/bioc/html/CAGEr.html).

Closest upstream antisense pair assignments

In order to define a set of 5'-GRO-seq cluster pairs that were reciprocally the closest upstream antisense of each other, a combination of BEDTools and custom scripts was used. BEDTools *closest* command (Quinlan and Hall, 2010) (settings: -S -id -D "a") was run on the modes of 5'-GRO-seq clusters (the position with the highest read count within a cluster) using the same file for both inputs. Custom Perl scripts were then used to parse the BEDTools output for only those cluster pairs where both modes were called as closest upstream antisense of each other.

DHS-based divergent and unidirectional promoter definitions

In order to define promoter DNase-I HyperSensitive regions (DHSs) as divergent or unidirectional, BEDTools (Quinlan and Hall, 2010) *intersect* command was used to find overlaps between DNaseI-seq peak calls (defining DHSs) and 5'-GRO-seq cluster modes, both described above. The output from BEDTools was then parsed with custom Perl scripts into different DHS categories. DHSs with exactly one intersecting TSS cluster mode were considered unidirectional. DHS with exactly two intersecting 5'-GRO-seq cluster modes where the two modes were upstream and antisense of each other, one annotating as TSS and the other as intergenic, were considered divergent. DHSs with more than one intersecting 5'-GRO-seq cluster modes on any one DNA strand, or with two 5'-GRO-seq cluster modes on opposite strands but downstream of each other, were removed from further analysis. For an increased-confidence unidirectional group, unidirectional classified DHSs intersecting reverse-side annotated TSSs (yet having no 5'GRO-seq clusters) or containing exactly one TSS-annotating cluster mode that was also part of the divergent or bidirectional reciprocal closest upstream antisense selection (described above) were considered ambiguous and removed from further analysis..

Heat map and meta-analysis plots

5'-GRO-seq and DNaseI-seq heat maps were made by calculating the center point between the 5'-GRO-seq cluster modes of the paired forward/reverse TSS clusters or the center point of DHSs. Windows were then taken around these center points and strand assignments (important for plotting orientation) made according to the forward, annotated, gene for divergent cluster pairs or unidirectional promoter DHS. For TSS-TSS or intergenic-intergenic 5'-GRO-seq cluster pairs, the cluster with higher read counts was used for strand assignment since there is no clear definition for sense/antisense in these situations. Genomic coordinates were then grouped in bins of 10 and the number of reads whose 5'end mapped to each bin were counted independent of strand and scaled so that the minimum value for each window is 0 and the maximum value is 1. Windows were sorted according to the distance between cluster pairs or the width of the DNaseI-seq peaks and plotted using the ggplot2 R package (Wickham, 2009).

For sequence heat maps, center positions, windows, strand and ranking were determined as above. BEDTools *getfasta* command (Quinlan and Hall, 2010) was used to retrieve the sequence corresponding to each window and ggplot2 (Wickham, 2009) was used for the plotting.

For TAF1 and Tbp ChIP-seq meta-analysis plots, sequence reads were extended by the fragment length calculated by SISSRS (Narlikar and Jothi, 2012). Center points, windows, and strands were determined as described above. For each window, genomic positions were grouped in bins of 10. If a bin overlapped a SISSRS summit (Narlikar and Jothi, 2012) (see above), then the number of extended-reads covering that bin were counted. If no peak summits overlapped a bin, it was assigned a 0. The per-bin means across all promoter locations were plotted using the ggplot2 R package (Wickham, 2009).

For GRO-seq metaplots center points, windows, and strands were determined as described above. For each window, genomic positions were grouped in bins of 10 and the number of sequence tag 5'ends counted per bin in a strand sensitive manner. The two resulting vectors of binned counts (one for each strand per window) were scaled together so that the minimum value for each window is 0 and the maximum value is 1. The per-bin means of these strand-sensitive, scaled, vectors across all promoter locations were plotted using the ggplot2 R package (Wickham, 2009).

The number of ENCODE CAGE sequence tag 5'ends were counted that intersected each 5'-GRO-seq cluster and the distribution of such counts per group were plotted as boxplots using the ggplot2 R package (Wickham, 2009).

Histone modification metaplot center points, windows, and strands were determined as described above. Reads were extended by the fragment sizes calculated within JAMM (Ibrahim et al., 2014). Genomic coordinates were then grouped in bins of 10 and the number of extended reads per million mapped overlapping each bin were counted independent of strand. The per-bin means across all promoter locations were plotted using the ggplot2 R package (Wickham, 2009).

TSS initiation pattern analysis

NarrowPeak, BroadPeak, and WeakPeak initiation patterns as defined previously(Ni et al., 2010) were determined for the specified groups from the 5'-GRO-seq clusters with at least 25 read counts.

Position-specific threemer counts

Position-specific threemer counts were determined with custom Perl scripts. After counting the instances of each threemer at each position, this value was divided by the total occurrence of that threemer in that sequence group. These values were plotted using the ggplot2 R package (Wickham, 2009).

Probabilistic model of transcription start sites

In order to compare the sequence composition of reverse direction core promoters and to scan DHS regions for transcription start site sequences, we employed a previously published position-specific Markov chain model (Frith et al., 2008) (PSMM). We used the first-order setting which will calculate the probability of a given di-nucleotide at a given position relative to that position's mono-nucleotide frequency, normalized for the di- and mono-nucleotide frequencies in the training set independent of position. Since the program reports log2 scores, all the values in our plots are 2^S, S being the log2 score output by the program.

A 10-fold cross validation scheme of the PSMM was implemented as follows. To train the model, the list of forward, TSS annotating, core promoters from either the closest upstream antisense selection, or DHS-based selection strategies, were split into 10 equal-size, non-overlapping, groups. These were designated as 10 unique "test" sets. For each test set, a corresponding training set was composed of the regions in the complete set that did not overlap the test set. The PSMM was then trained 10 separate times, once for each training set, on sequences +/- 50 bp around the TSS cluster modes. Each of the 10 models was then run on its corresponding, non-overlapping, test set. For the closest upstream antisense selection strategy, the test sequences were +/- 50 around the modes of the 5'-GRO-seq clusters. For the DHS-based

selection strategy, the test sequences were -150 to +50 around the appropriate DHS edge corresponding to the 5'-GRO-seq clusters of that test set.

In addition to the test group subsets, the 10 models were each run on the complete set of other sequences in question. For the closest upstream antisense selection strategy, these other sequences were +/- 50 around the 5'-GRO-seq cluster modes. Means were calculated for each position across the promoters of each list, resulting in 10 vectors of position means, one for each trained model. The mean at each position across these 10 vectors was plotted using ggplot2 (Wickham, 2009). For the DHS-based selection strategy, the sequences were -150 to + 50 around the appropriate DHS edge. Negative scores where the background model was higher than the TSS model were set to zero. The sequence positions were grouped in bins of 10 and the average score from each bin was calculated, then the mean average score was calculated for each binned position across all promoters of the list, resulting in 10 vectors of average score means at each position, one for each trained model. Shuffled control sequences were generated using the shuffleseq algorithm with default settings from the EMBOSS suite (Rice et al., 2000). The mean at each binned position across these 10 vectors was smoothed and plotted using ggplot2 (Wickham, 2009). For divergent pair scores in Supplemental Figure 3d, the scores for each 5-GRO-seq cluster mode were combined for each of the 10 cross validation runs and plotted as boxplots using ggplot2 (Wickham, 2009).

Receiver operator characteristic and precision recall curves were generated by defining true positives as the modes of 5'-GRO-seq clusters and true negatives as every other nucleotide in the tested windows, the results plotted for each of the 10 models from the closest upstream antisense selection using the R package ROCR (Sing et al., 2005).

Motif scanning

The TRANSFAC TATA-box binding protein or JASPAR Initiator position weight matrices (M00252; pwm) were used with the Scanner Toolset (Megraw et al., 2009) to scan sequences -35 to -25 upstream for TATA and +/- 5 for initiator around the forward TSS modes of the divergent and unidirectional promoter groups. A fixed first order Markov background was used for each list calculated from sequences +/- 50 around the forward TSS modes. Thresholds for fixed background scans were determined with a false positive rate cutoff of 0.001 as described in Megraw *et al.* (Megraw et al., 2009). For score distributions, highest scores were taken when locations contained multiple hits in the region scanned.

CpG island (CGI) analysis

Genomic coordinates of CGI were taken from the UCSC table browser (Kuhn et al., 2013). Either divergent or unidirectional DHSs were intersected with these coordinates using BEDTools intersect (Quinlan and Hall, 2010), either with the –u setting for counting the number of DHSs that intersect a CGI or the –wa –wb setting for determining size distributions of CGIs that intersect DHSs.

Chromatin State Segmentation

Similar to previous approaches (Ernst and Kellis, 2012; Hoffman et al., 2012), we employed a Hidden Markov Model (Taramasco and Bauer, 2013) (HMM) for unsupervised genome-wide clustering of histone modification ChIP-Seq read counts. We chose a multivariate Gaussian distribution for the HMM state emissions. Each chromatin state is a multivariate Gaussian distribution fully defined by its means vector, corresponding to the signals' means of the histone modification tracks, and its co-variance matrix.

In a pre-processing step, we define relevant locations for each histone modification (positions intersecting a ChIP-Seq peak) separately across the whole genome at 10-basepair resolution. The signal at relevant locations is defined as background-normalized, smoothed, extended-read counts (ie. ChIP-Seq signal). Peaks were identified using JAMM (Ibrahim et al., 2014), as described above. For each histone modification dataset, we extracted the corresponding ChIP-Seq signal for each peak at single-basepair resolution, using the SignalGenerator pipeline provided with JAMM (Ibrahim et al., 2014). JAMM's SignalGenerator output is then aligned to the genome in 10-basepair bins using the BEDOps (Neph et al., 2012) *bedmap* command (settings: --mean). Bins that did not intersect ChIP-Seq peaks are assigned a signal of zero. ChIP-Seq signal for each histone modification track is then scaled so that the minimum value is zero and the maximum value is 1000 and converted to log-space.

The resulting 10-basepair binned signal tracks for all histone modifications are matched up and bins that have a zero ChIP-Seq signal in all tracks are discarded. Bins that have a zero ChIP-Seq signal in one or more histone modification track(s) but not the other(s) are assigned a simulated normally-distributed background signal with a mean equal to the lowest bin signal value in the corresponding histone modification track and a variance of 0.1.

To learn the emission and transition parameters of the HMM, we employ the Baum-Welch algorithm (Bilmes, 1997; Taramasco and Bauer, 2013), initialized via k-means, on the signal tracks of chromosome 1. This learning process results in distinct chromatin states, each represented as a multivariate Gaussian distribution. The mean vector for each state defines the average ChIP-Seq signals of the histone modification tracks in the corresponding state. We 0-to-1 scale the means across each histone modification to define the prototypical chromatin states shown in Fig. 6a.

Finally, we employ the Viterbi decoding algorithm (Taramasco and Bauer, 2013; Viterbi, 1967) to assign a chromatin state to each 10-basepair bin in the genome that had a peak in at least one of the histone modification tracks. Locations that did not have a peak in any histone modification track (no relevant features, zero signal in all tracks) are not assigned a state. Book-ended bins that have the same state are merged. The output of this process is genome segmentation into variable-width non-overlapping chromatin states similar to Segway (Hoffman et al., 2012) and ChromHMM (Ernst and Kellis, 2012).

The main advantage of our chromatin state genome segmentation pipeline is that it allows for chromatin state assignment at high-resolution using "semi-binarized" signal, as opposed to using fully binarized (enriched / not-enriched) information at 200 bp resolution utilized in the ChromHMM approach (Ernst and Kellis, 2012). Our semi-binarized signal is the smoothed-extended ChIP-Seq read counts for relevant locations in the genome (ChIP-Seq peaks) and zeros elsewhere. This allows to account for information about the co-variance of the histone modifications' signals, but without suffering from noise over-representation, and thus has the potential to lead to more meaningful clustering of the histone modification signals compared to previous approaches(Ernst and Kellis, 2012; Hoffman et al., 2012). Finally, we do not analyze the entire genome, but only locations which had ChIP-Seq peaks in at least one histone modification dataset. Therefore, we can assign chromatin states at high-resolution 10 bp bins, close to the single-basepair resolution of Segway (Hoffman et al., 2012) but without its expensive computational resources requirement. Segway (Hoffman et al., 2012) can only run on high-performance computing clusters whereas our pipeline runs on typical desktop machines.

Chromatin State Analysis

To produce chromatin state coverage plots, we started with windows defined around the midpoints of DHSs as described above. Chromatin states were intersected with DHS-based windows using BEDTools (Quinlan and Hall, 2010) *intersect* command.

Chromatin state enrichment for different categories of 5'GRO cluster annotations were based on intersection of chromatin states with single-nucleotide locations that are 75-basepair downstream of the corresponding DHS edge.

Chromatin state enrichment with different genome-wide annotations were done using ChromHMM (Ernst and Kellis, 2012) *overlapEnrichment* command (settings: -b 10) using annotations based on hg19 UCSC knownGenes table (Kuhn et al., 2013) and hg19 Fantom5 enhancer list (Andersson et al., 2014).

Supplemental References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bilmes, J. (1997). A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. Tech. Rep., International Computer Science Insitute *ICSI-TR 97*.

Dodt, M., Roehr, J.T., Ahmed, R., and Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology *1*, 895–905.

Duttke, S.H.C. (2014). RNA Polymerase III Accurately Initiates Transcription from RNA Polymerase II Promoters in Vitro. The Journal of Biological Chemistry *289*, 20396–20404.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods *9*, 215–216.

Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. Genome Research *18*, 1–12.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods *9*, 473–476.

Ibrahim, M.M., Lacadie, S.A., and Ohler, U. (2014). JAMM: A Peak Finder for Joint Analysis of NGS Replicates. Bioinformatics (Oxford, England). doi: 10.1093/bioinformatics/btu568

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Briefings in Bioinformatics *14*, 144–161.

Lam, M.T.Y., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M., et al. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511–515.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Computational Biology *9*, e1003118.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078–2079.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal. http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., and Hatzigeorgiou, A.G. (2009). A transcription factor affinity-based code for mammalian transcription initiation. Genome Research *19*, 644–656.

Narlikar, L., and Jothi, R. (2012). ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. Methods in Molecular Biology (Clifton, N.J.) *802*, 305–322.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics (Oxford, England) *28*, 1919–1920.

Ni, T., Corcoran, D.L., Rach, E. a, Song, S., Spana, E.P., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nature Methods *7*, 521–527.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. Nature Structural & Molecular Biology.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841–842.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics : TIG *16*, 276–277.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics (Oxford, England) *21*, 3940–3941.

Taramasco, O., and Bauer, S. (2013). RHmm: Hidden Markov Models simulations and estimations. R package version 2.0.3 https://r-forge.r-project.org/R/?group_id=85.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory *13*, 260–269.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer Science & Business Media).